

Human Centred Object Co-Segmentation

Chenxia Wu¹, Jiemi Zhang², Ashutosh Saxena¹, and Silvio Savarese³

¹ Computer Science Department, Cornell University,
chenxiawu@cs.cornell.edu, asaxena@cs.cornell.edu

² Didi Research, jmzhang10@gmail.com

³ Computer Science Department, Stanford University, ssilvio@stanford.edu

Abstract. Co-segmentation is the automatic extraction of the common semantic regions given a set of images. Different from previous approaches mainly based on object visuals, in this paper, we propose a human centred object co-segmentation approach, which uses the human as another strong evidence. In order to discover the rich internal structure of the objects reflecting their human-object interactions and visual similarities, we propose an unsupervised fully connected CRF auto-encoder incorporating the rich object features and a novel human-object interaction representation. We propose an efficient learning and inference algorithm to allow the full connectivity of the CRF with the auto-encoder, that establishes pairwise relations on all pairs of the object proposals in the dataset. Moreover, the auto-encoder learns the parameters from the data itself rather than supervised learning or manually assigned parameters in the conventional CRF. In the extensive experiments on four datasets, we show that our approach is able to extract the common objects more accurately than the state-of-the-art co-segmentation algorithms.

1 Introduction

Image co-segmentation is defined as automatically extracting the common semantic regions, called *foregrounds*, given a set of images [30,16,4,14]. It provides an unsupervised way to mine and organize the main object segment from the image, which is useful in many applications such as object localization for assistive robotics, image data mining, visual summarization, *etc.*

The challenge of co-segmentation is that semantic labels are not given and we only have the visual information from the images. The only assumption is that the given images share some common semantic regions, *i.e.*, they belong to the same semantic category. So far, the dominating approaches are to co-discover these common regions only relying on their visual similarities [14,33,27,20]. For instance, [16,34,3] propose methods for unsupervised pixel-accurate segmentation of “similarly looking objects” in a given set of images. Vicente *et al.* [33] then introduce the concept of “objectness”, which follows the principle that the regions of interest should be “objects” such as bird or car, rather than “stuff” such as grass or sky. This helps focus the “attention” of the co-segmenter to discover common objects of interest as opposes to irrelevant regions.

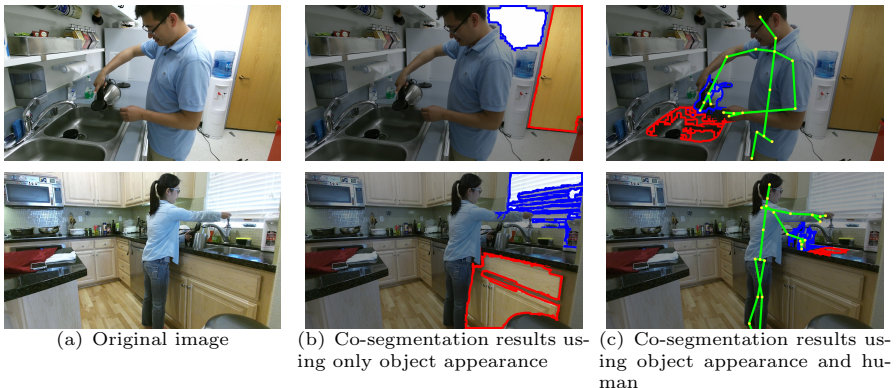


Fig. 1. We propose a human centred object co-segmentation approach by modeling both object visual appearance and human-object interactions. As in the example, human-object interactions help mining of more useful objects (the pots and the sinks) more accurately in the complex backgrounds, with view changes and occlusions. (Output foregrounds are blue, red circled and human skeletons are green colored.)

However, we argue that in situations where either objects’ appearance changes due to intra-class variations are severe (the sinks in Fig. 1) or when objects are observed under large view changes or partially occluded by other objects or humans (the pots in Fig. 1) or when multiple salient objects are present in the image (the door and the counter are also salient in Fig. 1), existing co-segmentation methods will not work well. In this paper, we argue that when images do contain humans that use objects in the scene (*e.g.*, a person opens a fridge, or washes dishes in a sink), which is typical in applications such as robotics, navigation or surveillance, *etc.*, we can leverage the interaction between humans and objects to help solve the co-segmentation problem. In essence, as a person interacts with an object in the scene, he/she provides an implicit cue that allows to identify the object’s spatial extend (*e.g.*, its segmentation mask) as well as the functional or affordances properties of the object (*i.e.*, the object regions that an human touches in order to use it).

Therefore, in this work, we propose a *human centred co-segmentation* method whereby the common objects are those often used by the observed humans in the scene and sharing the similar human-object interactions such as the pots and the sinks in Fig. 1-(c). We show that leveraging this information improves the results considerably compared to previous co-segmentation approaches.

The main challenge of a human centred object co-segmentation is to modeling the rich relations between objects as well as objects and humans in the image set. To achieve this, we first generate a set of object proposals as foreground candidates from the images. In order to discover the rich internal structure of these proposals reflecting their human-object interactions and visual similarities, we then leverage the power and flexibility of the fully connected conditional random field (CRF) [23] in a unsupervised setting and propose a *fully connected CRF auto-encoder*.

Our model uses the fully connected CRF to encode rich features to detect similar objects from the whole dataset. The similarity depends not only on object visual features but also a novel human-object interaction representation. We propose an efficient learning and inference algorithm to allow the full connectivity of the CRF with the auto-encoder, that establishes pairwise similarities on all pairs of the proposals in the dataset. As a result, the model selects the object proposals which have the most human interactions and are most similar to other objects in the dataset as the foregrounds. Moreover, the auto-encoder allows to learn the parameters from the data itself rather than supervised learning [17,18] or manually assigned parameters [14,15,36] as done in conventional CRF.

In the experiments, we show that our human centred object co-segmentation approach improves on the state-of-the-art co-segmentation algorithms on two human activity key frame Kinect datasets and a musical instrument RGB image dataset. To further show the generalization ability of the model, we also show a very encouraging co-segmentation result on a dataset combining the images without humans from the challenging Microsoft COCO dataset and the images with tracked humans.

In summary, the main contributions of this work are:

- We are the first to demonstrate that modeling human is useful to mining common objects more accurately in the unsupervised co-segmentation task.
- We propose an unsupervised fully connected CRF auto-encoder, and an efficient learning and inference approach to modeling rich relations between objects and humans.
- We show the leading performance of our human centred object co-segmentation in the extensive experiments on four datasets.

2 Related Work

Co-segmentation. Many efforts have been made on co-segmenting multiple images [6,19,26,2,25]. The early works used histogram matching [30], scribble guidance [4], or discriminative clustering [16] based on low-level descriptors to extract common foreground pixels. A mid-level representation using “objectness” was considered in [33,27] to extract similarly looking foreground objects rather than just common regions. Recently, Fu *et al.* [14] proposed an object-based co-segmentation from RGB-D images using the CRF models with mutex constraints. Our work also generates the object proposals using “objectness” and selects the foregrounds from the candidates. Differently, we are the first one to consider the human interaction to extract the foreground objects.

Early works on co-segmentation only considered two foreground and background classes. Recently, there are many co-segmentation methods which are able to handle multiple foreground objects. Kim *et al.* [21] proposed an anisotropic diffusion method by maximizing the overall temperature of image sites associated with a heat diffusion process. Joulin *et al.* [20] presented an effective energy-based method that combines a spectral-clustering term with a discriminative term, and an efficient expectation-minimization algorithm to optimize the

function. Lopamudra *et al.* [28] proposed a method by analyzing the subspace structure of related images. In [7,15], they presented a video co-segmentation method that extracts multiple foreground objects in a video set. Our work is also able to extract multiple foreground objects by formulating a fully connected CRF auto-encoder, which learns rich information from both objects and humans to extract the foregrounds.

Human-Object Interactions. Modeling human-object interactions or object affordances play an important role in recognizing both objects and human actions in previous works. The mutual context of objects and human poses were modeled in [38] to recognize human-object interactions that improve both human pose estimation and object detection. In [10,8,13], human-object interactions in still images or videos were modeled to improve action recognition. Wei *et al.* [35] modeled 4D human-object interactions to improve event and object recognition. In [17,18], hallucinated humans were added into the environments and human-object relations are modeled using the latent CRF to improve scene labeling. In this work, we show that human-object interactions provide an important evidence in the unsupervised mining of common objects from images. We also present a novel human-object interaction representation .

Learning Models. Our learning and inference model is extended from the work of CRF auto-encoder [1], which focuses on part-of-speech induction problems using a linear chain sequential latent structure with first-order Markov properties. However, the training and inference become impractical when using fully connected CRF with the auto-encoder. Thus, we introduce an efficient mean field approximation to make the computation still feasible. Our work is also close to the fully connected CRF [23] for the supervised semantic image segmentation, which also uses the mean field approximation to achieve an efficient inference. In contrast with this approach,, we use the mean field approximation to fast compute the gradients of two partition functions, which are exponential growing with object classes without the approximation in the CRF auto-encoder.

3 Problem Formulation

Our goal is to segment out common foreground objects from a given set of images. Similar to most co-segmentation approaches, we first generate a set of object proposals $X = \{x_i\}_{i=1,2,\dots,N}$ as foreground candidates from each image, where N is the total number of object proposals in the given set of images. Here we use selective search [32], which merges superpixels to generate proposals based on the hierarchical segmentation. This approach has been broadly used as the proposal method of choice by many state-of-the-art object detectors.

Let us assume there are K objects in the images. We then formulate the object co-segmentation as a probabilistic inference problem. We denote the object cluster assignment of x_i as $y_i \in \mathcal{Y} = [1, 2, \dots, K]$ and $Y = \{y_i\}_{i=1,2,\dots,N} \in \mathcal{Y}^N$. We want to infer the cluster assignment of a proposal using other proposals with the probability $p(y_i|X)$. We select the object proposals with the highest inference probabilities as the foregrounds, since they have the most human interactions and similar objects as determined by our probabilistic model.

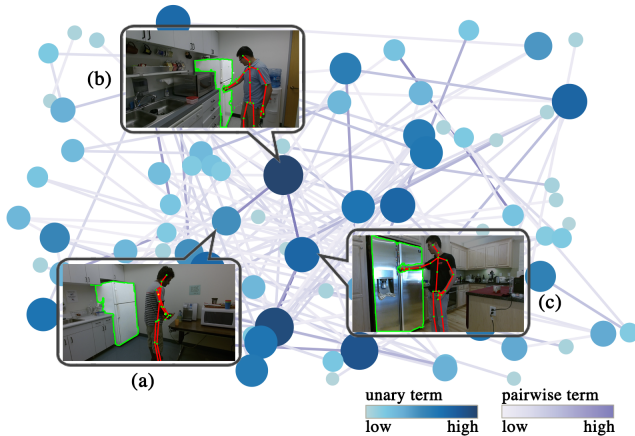


Fig. 2. Learned CRF graph from the data. The nodes are unary terms of object proposals encoding object appearance and human-object interaction features. The edges are pairwise terms encoding similarities on object appearance and human-object interactions. In the example, the fridges in (a) and (b) are visually similar, and the fridges in (b) and (c) have the similar human-object interactions. These similar objects with more human interactions are more likely to be segmented out in our approach, since they have higher unary terms and pairwise terms in the CRF graph. For a good visualization, we only plot the terms with respect to the most likely object cluster for each object proposal and the edges below a threshold are omitted.

4 Model Representation

We propose a fully connected CRF auto-encoder to discover the rich internal structure of the object proposals reflecting their human-object interactions and visual similarities. Unlike previous works relying mainly on the visual similarities between objects, we also encourage the objects with more human interactions and having more similar interactions with other objects to be segmented out. We plot a learned CRF graph from a set of images in Fig. 2. In the example, the fridges in (a) and (b) are visually similar, and the fridges in (b) and (c) are similar to each other on human-object interactions even though they look different. These similar objects with more human interactions are more likely to be segmented out in our approach, since they have higher terms in the CRF graph.

The fully connected CRF auto-encoder consists of two parts (The graphic model is shown in Fig. 3). The encoding part is modeled as a fully connected CRF, which encodes the observations x_i of the object proposal into object cluster hidden nodes y_i . The reconstruction part reconstructs the hidden nodes by generating a copy of the observation itself \hat{x}_i , which considers that a good hidden structure should permit reconstruction of the data with high probability [1],

4.1 Fully Connected CRF Encoding

We first introduce how the fully connected CRF encodes the observations of the object proposals. In CRF, the conditional distribution is given by $P_\lambda(Y|X) =$

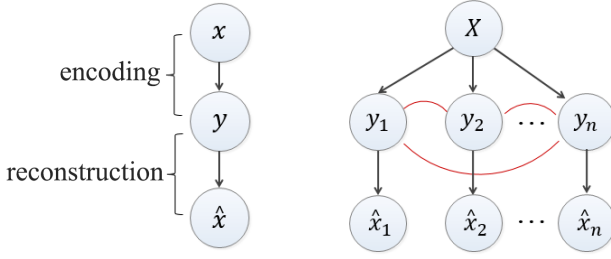


Fig. 3. Graphic model of our fully connected CRF auto-encoder.

$\frac{1}{Z} \exp\{\Phi_\lambda(X, Y)\}$, where $Z = \sum_{Y' \in \mathcal{Y}^N} \exp\{\Phi_\lambda(X, Y')\}$ is a partition function and $\Phi_\lambda(\cdot)$ is called the energy function defined as follows:

$$\Phi_\lambda(X, Y) = \underbrace{\sum_i \lambda^{(u)}(y_i)^\top x_i}_{\text{unary terms}} + \underbrace{\sum_{i < j} \lambda^{(p)}(y_i, y_j) S(x_i, x_j)}_{\text{pairwise terms}}, \quad (1)$$

where $\lambda^{(u)}(y_i)^\top x_i$ is the unary term that encodes object visual appearance features and human-object interaction features:

$$\lambda^{(u)}(y_i)^\top x_i = \underbrace{\lambda^{(uo)}(y_i)^\top f_i}_{\text{object appearance}} + \underbrace{\lambda^{(uh)}(y_i)^\top h_i}_{\text{human-object interaction}}. \quad (2)$$

Object Appearance Modeling. In Eq. (2), $\lambda^{(uo)}(y_i)^\top f_i$ encodes the object visual feature vector f_i by the linear weights $\lambda^{(uo)}(k)$ of each cluster k . It encourages an object to give larger value $\lambda^{(uo)}(k)^\top f_i$ if it belongs to the object cluster k . We use rich kernel descriptors by kernel principal component analysis [5] on the input images: gradient, color, local binary pattern for the RGB image, depth gradient of depth image and spin, surface normals of point cloud for the Kinect data, which have been proven to be useful features for scene labeling [29,36].

Human-Object Interaction Modeling. In Eq. (2), $\lambda^{(uh)}(y_i)^\top h_i$ encodes the human-object interaction feature vector h_i by the linear weights $\lambda^{(uh)}(k)$ of each cluster k .

To capture the interactions between objects and humans, we propose a novel feature to represent physical human-object interactions such as sitting on the chair, opening the fridge, using their spatial relationships. This feature helps detect those objects used by humans in the scene.

We illustrate the feature representation in Fig. 4 for RGB-D data. In detail, we convert the depth image into the real-world 3D point cloud and are also given the 3D coordinate of each joint of a tracked human. Each human body part is represented as a vector starting from a joint to its neighboring joint (Fig. 4-(a)(b)). Then we consider a cylinder with the body part as the axis and divide the cylinder into 15 bins by segmenting the body part vertically into 3 parts and the circle surrounding the body part into 5 regions evenly⁴ (Fig. 4-(b)). Given

⁴ To avoid the affect of points of the human part, we do not consider the innermost circle.

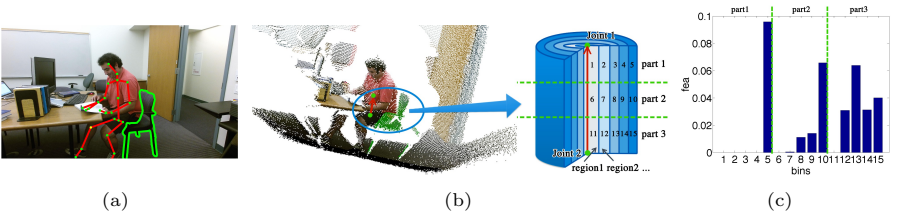


Fig. 4. Our human-object interaction feature for RGB-D data. In (a), we are given the joints (green dots) of the tracked human, and a object proposal region (green mask). In (b), we divide the cylinder surrounding each body part vector (red line, *spine-base* to *spine-mid* in the example) into 15 bins by segmenting the body part vertically into 3 parts and the circle surrounding the body part into 5 regions. In (c), we compute the histogram of the points in these 15 bins and normalize it by the total number of the total points in the object proposal.

the point cloud in an object proposal region, we calculate the histogram of the points in these 15 bins and normalize it by the number of the total points in the object proposal as the final feature h_i (Fig. 4-(c)). For multiple people in the scene, we compute the max of the histograms of all humans.

For RGB only data, we assume a 2D bounding box of human is given by a person detector. We divide the bounding box evenly into 6×6 bins and compute the normalized histogram of pixels within the bins.

The feature captures the distributions of the object points relative to the human body. It can represent the relative position, size between the humans and objects as well as the human poses and object shapes especially in 3D space. For example in Fig. 4-(c), we plot the histogram feature of the body part *spine-base* to *spine-mid* relative to the *chair*. We can see that the distribution reflects the shape of the chair in the following way: the points of the chair-back lie in the distant bins of the upper part (mostly in bin 5 among bin 1-5), and from the chair-back to the chair-base, points become more evenly distributed from the middle part (bin 6-10) to the lower part (bin 11-15).

Since our feature vector h_i has larger values for more human-object interactions, we constrain the weights of the human-object interaction features $\lambda^{(uh)}(k) \geq 0$ to encourage more interactions.

Pairwise Similarity Modeling. In Eq. (1), $\lambda^{(p)}(y_i, y_j)S(x_i, x_j)$ is the pairwise term that encodes similarities on object visual appearance and human-object interactions of each object pair:

$$\lambda^{(p)}(y_i, y_j)S(x_i, x_j) = \underbrace{\lambda^{(po)}(y_i, y_j)S(f_i, f_j)}_{\text{object similarities}} + \underbrace{\lambda^{(ph)}(y_i, y_j)S(h_i, h_j)}_{\text{interaction similarities}}, \quad (3)$$

where $S(a, b) = \exp\{\frac{-\|a-b\|_2^2}{\delta}\}$ is the similarity function between features⁵. Note that we consider same class to be similar to each other, so we constrain $\lambda^{(p)}(k, k) > 0$ and $\lambda^p(k, l) = 0, k \neq l$.

4.2 Reconstruction

To independently generate \hat{x} given y , we define the reconstruction model as a multivariate normal distribution:

$$P_\theta(\hat{x}|y) = N(\hat{x}|\theta^{(\mu)}(y), \theta^{(\Sigma)}(y)), \quad (4)$$

where $\theta^{(\mu)}(k), \theta^{(\Sigma)}(k)$ are the mean and covariance of the normal distributions of each class k .

5 Model Learning and Inference

In this section, we introduce how we learn the parameters in our fully connected CRF autoencoder from the data and the inference.

Eq. 5 gives the parametric model for the observations X :

$$\begin{aligned} P_{\lambda, \theta}(\hat{X}|X) &= \sum_{Y \in \mathcal{Y}^N} \overbrace{P_\lambda(Y|X)}^{\text{encoding}} \overbrace{P_\theta(\hat{X}|Y)}^{\text{reconstruction}} \\ &= \sum_{Y \in \mathcal{Y}^N} \frac{\exp\{\Phi_\lambda(X, Y)\}}{\sum_{Y' \in \mathcal{Y}^N} \exp\{\Phi_\lambda(X, Y')\}} \prod_i P_\theta(\hat{x}_i|y_i) \\ &= \frac{\sum_{Y \in \mathcal{Y}^N} \exp\{\Phi_\lambda(X, Y) + \log \sum_i P_\theta(\hat{x}_i|y_i)\}}{\sum_{Y' \in \mathcal{Y}^N} \exp\{\Phi_\lambda(X, Y')\}}, \end{aligned} \quad (5)$$

where λ, θ are the parameters of the encoding parts $P_\lambda(Y|X)$ and the reconstruction parts $P_\theta(\hat{X}|Y)$.

5.1 Efficient Learning and Inference

We maximize the regularized conditional log likelihood of $P_{\lambda, \theta}(\hat{X}|X)$ over all the object proposals to learn the parameters λ, θ :

$$\log L(\lambda, \theta) = R_1(\lambda) + R_2(\theta) + \log \sum_Y P_\lambda(Y|X) P_\theta(\hat{X}|Y), \quad (6)$$

where $R_1(\lambda), R_2(\theta)$ are the regularizers. Note that the space of Y is $[1, 2, \dots, N]^K$, which is exponential to object cluster number. Maximizing the above function requires computing gradients, while the summation over Y, Y' in the space is intractable.

⁵ In practice, we use the augmented features $f'_i = [f_i, -1], h'_i = [h_i, -1], S'(x_i, x_j) = [S(x_i, x_j), -1]$ to also learn the bias of the features. So $\lambda^{(uo)}(k) = [\omega^{(uo)}, b^{(uo)}], \lambda^{(uh)}(k) = [\omega^{(uh)}, b^{(uh)}], \lambda^{(p)}(k, k) = [\omega^{(p)}, b^{(p)}]$ also add another dimension, where ω are the importance weights and $b > 0$ is the bias.

Mean Field Approximation. In order to make the learning and inference efficient, we use a mean field approximation that has been widely used in the complex graph inference [11,23]. To use mean field approximation in our model, we first introduce two probabilities of Y :

$$P_{\lambda,\theta}(Y) = \frac{1}{Z} \exp\{\Phi_\lambda(X, Y) + \log P_\theta(\hat{X}|Y)\}, \quad P'_\lambda(Y) = \frac{1}{Z'} \exp\{\Phi_\lambda(X, Y)\}, \quad (7)$$

where

$$Z = \sum_{Y \in \mathcal{Y}^n} \exp\{\Phi_\lambda(X, Y) + \log P_\theta(\hat{X}|Y)\}, \quad Z' = \sum_{Y \in \mathcal{Y}^n} \exp\{\Phi_\lambda(X, Y)\}, \quad (8)$$

are the partition functions summing over Y to make two probabilities valid. Then Eq.(6) can be rewritten as:

$$\log L(\lambda, \theta) = R_1(\lambda) + R_2(\theta) + \log Z - \log Z'. \quad (9)$$

Instead of directly computing $P_{\lambda,\theta}(Y), P'_\lambda(Y)$, we use the mean field approximation to compute the distributions $Q(Y), Q'(Y)$ that minimize the KL-divergence $D(Q||P_{\lambda,\theta}), D(Q'||P'_\lambda)$. Then all distributions Q, Q' can be expressed as a product of independent marginals, $Q(Y) = \prod_i Q_i(y_i), Q'(Y) = \prod_i Q'_i(y_i)$ [11]. As a result, the gradients of $\log Z$ and $\log Z'$ can be easily computed approximately (see Eq. (11)).

By minimizing the KL-divergence, $Q_i(Y), Q'_i(Y)$ has the following iterative update equations:

$$\begin{aligned} Q_i(y_i = k) &= \frac{1}{Z_i} \exp\{\lambda^{(u)}(k)^\top x_i + \lambda^{(p)}(k, k) \sum_{j \neq i} S(x_i, x_j) Q_j(k) \\ &\quad + \log N(\hat{x}_i | \theta^{(\mu)}(k), \theta^{(\Sigma)}(k))\}, \\ Q'_i(y_i = k) &= \frac{1}{Z'_i} \exp\{\lambda^{(u)}(k)^\top x_i + \lambda^{(p)}(k, k) \sum_{j \neq i} S(x_i, x_j) Q'_j(k)\}. \end{aligned} \quad (10)$$

The detailed derivation is in the supplementary material. Following the above equations, the computation of Q, Q' can be done using an iterative update as shown in Algorithm 1.

In each update, it costs $O(N^2 \times K)$. It also can be used a similar message passing by the high-dimensional filtering as in [23] to speed up the update, which is not the focus of the paper. The update was mostly converged within 10 rounds in our experiments (see Fig. 5(b)).

Learning. We iteratively learn λ and θ by maximizing the log likelihood with respect to each type of parameters. We update λ using the Adagrad approach [12] which computes the gradients and update θ using EM [9]. Using the mean field approximation described above, we can easily compute the gradients of $\log Z, \log Z'$ in Eq.(9), leading to a simple approximation of the gradients:

$$\begin{aligned} \frac{\partial(\log Z - \log Z')}{\partial \lambda^{(u)}(k)} &= \sum_i x_i (Q_i(k) - Q'_i(k)), \\ \frac{\partial(\log Z - \log Z')}{\partial \lambda^{(p)}(k, k)} &= \sum_{i \neq j} S(x_i, x_j) (Q_i(k) Q_j(k) - Q'_i(k) Q'_j(k)). \end{aligned} \quad (11)$$

Algorithm 1 Mean field to compute Q and Q' .

Initialize Q and Q' :

$$Q_i(k) = \frac{1}{Z_i} \exp\{\lambda^{(u)}(k)^\top x_i + \log N(\hat{x}_i | \theta^{(\mu)}(k), \theta^{(\Sigma)}(k))\},$$

$$Q'_i(k) = \frac{1}{Z'_i} \exp\{\lambda^{(u)}(k)^\top x_i\},$$

while not converged **do**

$$\hat{Q}_i(k) = \lambda^{(p)}(k, k) \sum_{j \neq i} S(x_i, x_j) Q_j(k),$$

$$Q_i(k) = \exp\{\lambda^{(u)}(k)^\top x_i + \hat{Q}_i(k) + \log N(\hat{x}_i | \theta^{(\mu)}(k), \theta^{(\Sigma)}(k))\},$$

$$\hat{Q}'_i(k) = \lambda^{(p)}(k, k) \sum_{j \neq i} S(x_i, x_j) Q'_j(k),$$

$$Q'_i(k) = \exp\{\lambda^{(u)}(k)^\top x_i + \hat{Q}'_i(k)\},$$

$$\text{normalize } Q_i(k), Q'_i(k).$$

end while

The detailed derivation is in the supplementary material.

Inference. After learning the parameters λ, θ , we can infer the posterior, conditioning on both observations and reconstructions, $\hat{Y} = \arg \max_Y P_{\lambda, \theta}(Y|X, \hat{X})$. This is proportional to $P_{\lambda, \theta}(Y)$ in Eq.(7), which can be efficiently computed using the approximation probability $Q(Y) = \prod_i Q_i(y_i)$. So the probability of each object proposal for each class $p_{\lambda, \theta}(y_i|X, \hat{X}) \propto Q_i(y_i)$, which we use as the confidence score to extract the foreground objects of each image.

6 Experiments

6.1 Compared Baselines

We compared our human centred object co-segmentation approach with two state-of-the-art algorithms: a multi-class image co-segmentation approach by combining spectral- and discriminative-clustering (Joulin *et al.* 2012 [20]) and a CRF based solution with the manually assigned parameters (Fu *et al.* 2014 [15]). Fu *et al.* 2014 [15] is designed for co-segmenting foregrounds from videos, so we keep their features for the static frame and remove the temporal features. We run their source code on the given RGB images with the default settings. We also evaluate our model using object-only visual features in the experiments.

6.2 Evaluations

We use the same evaluation metrics as in multi-class co-segmentation [20]: the intersection-over-union score defined as $\max_k \frac{1}{|\mathbf{I}|} \sum_{i \in \mathbf{I}} \frac{GT_i \cap R_i^k}{GT_i \cup R_i^k}$, where \mathbf{I} is the image dataset, GT_i is the ground-truth region and R_i^k is the region associated with the k -th cluster in image i . The evaluation validates that the foreground object is indeed rather well represented by one of the object clusters in most test cases, which may be sufficient to act as a filter in most applications such as [31]. As discussed in multi-class co-segmentation [20], we set K to be more than the number of groundtruth foreground object classes plus background class to give better segmentation results. We set $K = 4$ for the data with a single foreground and $K = 6$ for the data with two foregrounds. We also evaluate how performance varies with K in the experiments.

6.3 Datasets

Human Activity Key Frame Kinect Dataset We first evaluate on human activity key frames extracted from the activity videos in two datasets: Cornell Activity Dataset-120 (CAD-120)⁶ [22] recorded by Microsoft Kinect v1, and Watch-n-Patch⁷ [37] recorded by the Microsoft Kinect v2. Each frame in the datasets has a registered RGB and depth image pair as well as tracked human skeletons.

CAD-120 contains activity sequences of ten different high level activities performed by four different subjects. We evaluate the image co-segmentation with the top four existing foreground objects in CAD-120: microwave, bowl, box, cup. For each type, we extract the 60 key frames of the activity videos containing the object. We evaluate the foreground regions using the provided groundtruth bounding box.

Watch-n-Patch activity dataset contains human daily activity videos performed by 7 subjects in 8 offices and 5 kitchens with complex backgrounds. In each environment, the activities are recorded in different views. In each video, one person performed a sequence of actions interacting with different types of objects. We evaluate three types of scenes in the dataset, each of which has two foreground objects: table and chair, fridge and microwave, pod and sink. For each scene, we extract the 70 key frames of the relevant activity video containing the objects. We label the groundtruth foreground pixels for evaluation.

People Playing Musical Instrument Dataset. We also evaluate on a RGB dataset to see the performance using RGB only features. We use the People Playing Musical Instrument (PPMI) dataset⁸, which was used to evaluate recognizing human-object interactions in [38]. It contains RGB images of people ‘playing’ or ‘with’ different types of musical instruments. Some of the images have multiple people interacting with the instruments. We evaluate three types of instruments, cello, French horn and violin by randomly selecting 80 images from each class and label the foreground pixels in the image for evaluation.

MS COCO combining with Watch-n-Patch Dataset Since image data with humans are not always available, we finally give the co-segmentation results on the images without humans from the challenging Microsoft COCO (MS COCO) dataset [24] combining with a small portion of images with tracked humans from Watch-n-Patch dataset. The images from MS COCO dataset has more clustered backgrounds from variant sources. We evaluate three classes: chair, fridge and microwave. For each class, we randomly select 50 images from indoor scenes in MS COCO dataset and 20 images from Watch-n-Patch dataset, then combine them as the test set. We use the same RGB features and human features as described above.

Note that one challenge for image co-segmentation is the scalability, as the state-of-the-art algorithms rely on heavy computations on relation graphs. Therefore, most evaluation datasets in previous works [19,33,20,14] have less than 50

⁶ <http://pr.cs.cornell.edu/humanactivities/data.php>

⁷ <http://watchnpatch.cs.cornell.edu/>

⁸ <http://ai.stanford.edu/~bangpeng/ppmi.html>

Table 1. Co-Segmentation results on CAD-120 dataset (%).

class	microwave	bowl	box	cup
Joulin <i>et al.</i> 2012 [20]	21.6	22.5	19.2	17.7
Fu <i>et al.</i> 2014 [15]	47.5	14.9	40.2	9.3
Ours (object-only)	45.1	19.7	32.6	22.8
Ours	54.3	24.8	38.2	27.9

Table 2. Co-Segmentation results on Watch-n-Patch dataset (%).

class	table	chair	fridge	microwave	pod	sink
Joulin <i>et al.</i> 2012 [20]	34.7	17.2	29.9	5.5	5.3	17.9
Fu <i>et al.</i> 2014 [15]	21.6	15.7	25.4	24.5	21.3	23.6
Ours (object-only)	41.9	26.9	33.1	17.5	20.4	23.2
Ours	50.0	36.4	44.7	20.5	23.7	28.6

images per class. In our experiments, the test set has more than 50 but still less than 100 images per class.

6.4 Results

We give the results in Table 1, 2, 3, 4. We can see that in most cases, our approach performs better than the state-of-the-art image co-segmentation methods. We discuss our results in the light of the following questions.

Did modeling human-object interaction help? In most cases, we can see that our approach to modeling the human-object interaction gives the best result. This is because more human-object interactions give the higher unary term for the interesting objects interacting with the humans, and the similar human-object interactions link the same objects in the CRF graph with larger pairwise terms even though they may not look similar. As a result, we are able to segment out the common object accurately even with view, scale changes, occlusions and in complex backgrounds. We show some example results in Fig. 6 and Fig. 7.

How successful is our fully connected CRF auto-encoder? From the results, we can see that our fully connected CRF auto-encoder model using the object only features also performs better than other algorithms. This is because our model is able to learn the parameters from the data itself rather than manually assigning parameters of the typical CRF model in Fu *et al.* 2014 [15], then the model is more data dependent and does not require much parameter tuning. Though Fu *et al.* 2014 [15] performed well in some cases, the approach is not stable as the parameters are preset. Also, benefit from our efficient learning and inference algorithm, we are able to use fully connected hidden nodes to model the rich relations between all objects and humans in the dataset, so that we have more information to detect the common foreground objects.

Can human information be generalized to the segments without humans? In our first three datasets with humans in the image, there are a few images where humans are not interacting with the foreground objects such as the fridge on the right in Fig. 6(b). In these cases, our approach was still possible to segment it out correctly, since it was linked to other visually similar objects which are interacting with humans in the fully connected CRF graph.

Table 3. Co-Segmentation results on PPMI dataset (%).

class	cello	frenchhorn	violin
Joulin <i>et al.</i> 2012 [20]	21.9	20.0	18.1
Fu <i>et al.</i> 2014 [15]	30.1	40.8	26.4
Ours (object-only)	34.5	41.0	28.3
Ours	36.2	49.2	31.5

Table 4. Co-Segmentation results on MS COCO + Watch-n-Patch dataset (%).

class	chair	fridge	microwave
Joulin <i>et al.</i> 2012 [20]	4.2	14.1	10.3
Fu <i>et al.</i> 2014 [15]	6.9	11.4	9.2
Ours (object-only)	7.5	10.2	10.5
Ours	12.5	17.5	15.9

From the results in Table 4 on the MS COCO + Watch-n-Patch dataset, we can see that the task is very challenging since images are from different domains with more clustered backgrounds. Though all compared methods perform not more than 20 percent in accuracy, modeling humans even in a few images still improves the performance.

2D bounding box of human vs. 3D human skeleton. From Table 3, we can see that even in 2D images with the bounding box of the detected human, modeling the human-object interactions improves the co-segmentation performance. Using our proposed 3D human-object representation on the more accurate tracked humans gives more improvements as shown in Table 1, 2. In the examples in Fig. 6(b), we also found that our 3D human-object representation is useful to deal with the occlusions and view changes, which is challenging for object visual appearance only based co-segmentation approaches.

How performance varies with cluster number K ? We show the performance varying with the cluster number K in Fig. 5(a). We can see that the accuracy increase with the class number K as it has higher chance to hit the ground truth regions and more backgrounds are modeled. Our approach has the best performance for each K .

How fast is the learning? We also plot a learning curves of our model in Fig. 5(b). The learning of our approach can be converged mostly within 10 iterations.

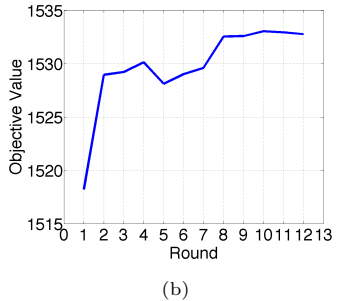
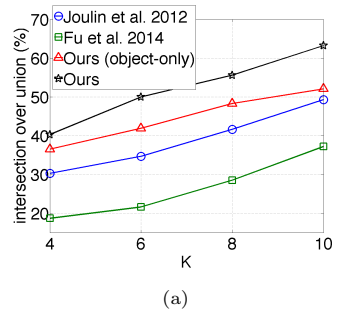


Fig. 5. (a). Results of table class on Watch-n-Patch dataset varying with cluster number K . (b). Learning curve of our approach.

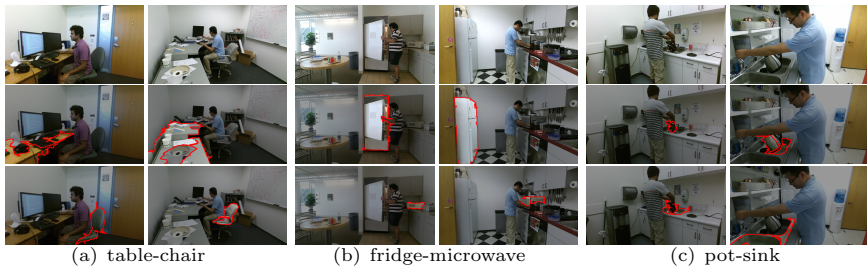


Fig. 6. Visual examples of our co-segmentation results on Watch-n-Patch dataset.

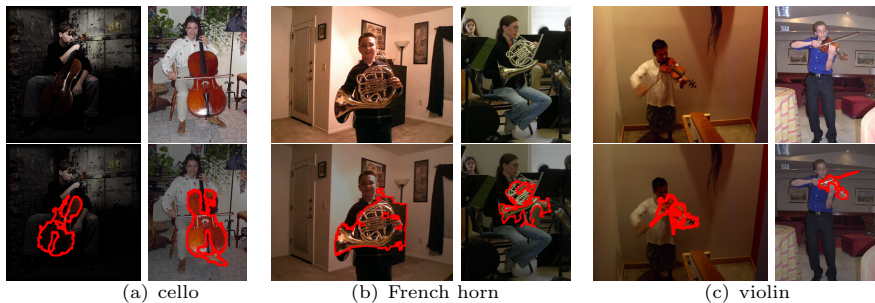


Fig. 7. Visual examples of our co-segmentation results on PPMI dataset.

7 Conclusion and Future Work

In this work, we proposed a novel human centered object co-segmentation approach using a fully connected CRF auto-encoder. We encoded a novel human-object interaction representation and rich object visual features as well as their similarities using the powerful fully connected CRF. Then we used the auto-encoder to learn the parameters from the data itself using an efficient learning even for this complex structure. As a result, we were able to extract those objects which have the most human interactions and are most similar to other objects in the dataset as the foregrounds. In the experiments, we showed that our approach extracted foreground objects more accurately than the state-of-the-art algorithms on two human activity Kinect key frame dataset as well as the musical instruments RGB image dataset. We also showed that our model was able to use the human information in a small portion of images to improve the co-segmentation results.

In the future, we consider extending our human centered object co-segmentation approach into the semi-supervised setting and incorporating temporal information for video data.

References

1. Ammar, W., Dyer, C., Smith, N.A.: Conditional random field autoencoders for unsupervised structured prediction. In: *Advances in Neural Information Processing Systems (NIPS)* (2014)
2. Baraldi, L., Grana, C., Cucchiara, R.: Scene segmentation using temporal clustering for accessing and re-using broadcast video. In: *Multimedia and Expo (ICME), IEEE International Conference on* (2015)
3. Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T.: icoseg: Interactive co-segmentation with intelligent scribble guidance. In: *CVPR* (2010)
4. Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T.: Interactively co-segmentating topically related images with intelligent scribble guidance. *International Journal of Computer Vision (IJCV)* 93(3), 273–292 (2011)
5. Bo, L., Ren, X., Fox, D.: Kernel Descriptors for Visual Recognition. In: *Advances in Neural Information Processing Systems (NIPS)* (2010)
6. Brostow, G., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: *ECCV* (2008)
7. Chiu, W.C., Fritz, M.: Multi-class video co-segmentation with a generative multi-video model. In: *CVPR* (2013)
8. Delaitre, V., Sivic, J., Laptev, I.: Learning person-object interactions for action recognition in still images. In: *Advances in Neural Information Processing Systems (NIPS)* (2011)
9. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* pp. 1–38 (1977)
10. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on* (2010)
11. D.Koller, N.Friedman.: *Probabilistic graphical models: Principles and techniques*. In: MITPress (2009)
12. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res. (JMLR)* 12, 2121–2159 (2011)
13. Escorcia, V., Niebles, J.: Spatio-temporal human-object interactions for action recognition in videos. In: *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on* (2013)
14. Fu, H., Xu, D., Lin, S., Liu, J.: Object-based rgb-d image co-segmentation with mutex constraint. In: *CVPR* (2015)
15. Fu, H., Xu, D., Zhang, B., Lin, S.: Object-based multiple foreground video co-segmentation. In: *CVPR* (2014)
16. Hochbaum, D., Singh, V.: An efficient algorithm for co-segmentation. In: *ICCV* (2009)
17. Jiang, Y., Koppula, H., Saxena, A.: Hallucinated humans as the hidden context for labeling 3d scenes. In: *CVPR* (2013)
18. Jiang, Y., Saxena, A.: Infinite latent conditional random fields for modeling environments through humans. In: *Robotics: Science and Systems (RSS)* (2013)
19. Joulin, A., Bach, F., Ponce, J.: Discriminative clustering for image co-segmentation. In: *CVPR* (2010)
20. Joulin, A., Bach, F., Ponce, J.: Multi-class cosegmentation. In: *CVPR* (2012)
21. Kim, G., Xing, E.P., Fei-Fei, L., Kanade, T.: Distributed Cosegmentation via Sub-modular Optimization on Anisotropic Diffusion. In: *ICCV* (2011)

22. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. *Int. J. Rob. Res.* 32(8), 951–970 (2013)
23. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: *Advances in Neural Information Processing Systems (NIPS)* (2011)
24. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. *CoRR* abs/1405.0312 (2014)
25. Lucchi, A., Marquez-Neila, P., Becker, C., Li, Y., Smith, K., Knott, G., Fua, P.: Learning structured models for segmentation of 2-d and 3-d imagery. *Medical Imaging*, *IEEE Transactions on* 34(5), 1096–1110 (2015)
26. Manfredi, M., Grana, C., Cucchiara, R.: Learning superpixel relations for supervised image segmentation. In: *International Conference on Image Processing (ICIP)* (2014)
27. Meng, F., Li, H., Liu, G., Ngan, K.N.: Object co-segmentation based on shortest path algorithm and saliency model. *Multimedia*, *IEEE Transactions on* (TMM) 14(5), 1429–1441 (2012)
28. Mukherjee, L., Singh, V., Xu, J., Collins, M.: Analyzing the subspace structure of related images: Concurrent segmentation of image sets. In: *ECCV* (2012)
29. Ren, X., Bo, L., Fox, D.: RGB-(D) scene labeling: Features and algorithms. In: *CVPR* (2012)
30. Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In: *CVPR* (2006)
31. Russell, B., Freeman, W., Efros, A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: *CVPR* (2006)
32. Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. *International Journal of Computer Vision (IJCV)* 104(2), 154–171 (2013)
33. Vicente, S., Rother, C., Kolmogorov, V.: Object cosegmentation. In: *CVPR* (2011)
34. Vicente, S., Kolmogorov, V., Rother, C.: Cosegmentation revisited: Models and optimization. In: *ECCV* (2010)
35. Wei, P., Zhao, Y., Zheng, N., Zhu, S.: Modeling 4d human-object interactions for event and object recognition. In: *ICCV* (2013)
36. Wu, C., Lenz, I., Saxena, A.: Hierarchical semantic labeling for task-relevant rgb-d perception. In: *Robotics: Science and Systems (RSS)* (2014)
37. Wu, C., Zhang, J., Savarese, S., Saxena, A.: Watch-n-patch: Unsupervised understanding of actions and relations. In: *CVPR* (2015)
38. Yao, B., Fei-Fei, L.: Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* 34(9), 1691–1703 (2012)